

## CSCI1950-L Midterm

*Due: Wednesday, March 19, 2008 at 3:00 pm*

*To be given to Prof. Istrail's assistant Erin Klopfenstein (CIT 546, ehk@cs.brown.edu)*

*Late submissions will not be accepted.*

**IMPORTANT:** The Exam has 5 Sections. You **MUST** do the required problems in Sections 1-4. Three problems in Section 1 are labeled **OPTIONAL**. You must choose two (2) of the three to answer. Extra work will be considered for extra credit.

### Contents

<b>1 ALIGNMENT (60)</b>	<b>2</b>
1.1 Dynamic Programming - Edit Graph (5) . . . . .	2
1.2 Role of 0 in Local Alignment (5) . . . . .	2
1.3 Global Alignment of 3 Sequences (OPTIONAL 15) . . . . .	2
1.4 Affine Gaps (5) . . . . .	2
1.5 Traceback (10) . . . . .	2
1.6 Endosymbiont Problem (OPTIONAL 15) . . . . .	3
1.7 Classic Alignments (5) . . . . .	3
1.8 Protein Structure (OPTIONAL 15) . . . . .	3
<b>2 HOW NOT TO LIE WITH STATISTICS (10)</b>	<b>5</b>
<b>3 SUBSTITUTIONS MATRICES (15)</b>	<b>6</b>
3.1 BLOSUM-like Matrices (10) . . . . .	6
3.2 Dayhoff Matrices (5) . . . . .	7
<b>4 GENOME ASSEMBLY (15)</b>	<b>8</b>
4.1 Consensus Sequence (10) . . . . .	8
4.2 Assembly Algorithms (5) . . . . .	8
<b>5 EXTRA CREDIT PROBLEMS</b>	<b>9</b>
5.1 "The Adventure of the Dancing Men" . . . . .	9
5.2 Tandem Repeats . . . . .	9

## 1 ALIGNMENT (60)

### 1.1 Dynamic Programming - Edit Graph (5)

Draw the Edit Graph for the global alignment of the following two sequences using the unitary similarity matrix:

$$\begin{array}{rcl} X & = & C \ C \ T \ A \ C \\ Y & = & A \ C \ C \ C \end{array}$$

Make sure you label clearly:

- the two sequences
- edge types
- coordinates
- direction of alignment
- start and end points
- scores
- optimal path(s)

After drawing the graph, identify the optimal path(s).

### 1.2 Role of 0 in Local Alignment (5)

Explain the role of 0 in the the Smith-Waterman algorithm. How come such a simple change in the global alignment algorithm could transform the algorithm to compute local alignments.

### 1.3 Global Alignment of 3 Sequences (OPTIONAL 15)

Generalize the recurrence formula for the global alignment algorithm for two DNA sequences to an algorithm for aligning three DNA sequences. What is the complexity of the new algorithm?

### 1.4 Affine Gaps (5)

Define what an affine gap scoring function is, and explain why the complexity of the local alignment algorithm with affine gap scoring function for two sequences of sizes  $M$  and  $N$  is  $O(MN)$ .

### 1.5 Traceback (10)

What traceback algorithm would you use in a global alignment of two sequences if you want to retrieve all the optimal alignments as well as suboptimal alignments with cost one less than optimal?

## 1.6 Endosymbiont Problem (OPTIONAL 15)

*Wolbachia pipientis* is an endosymbiont that infects a wide range of insects. This bacteria is often maternally inherited, because it can infect germ-line cells. Recently, gene transfer has been reported between *Wolbachia* and one of its hosts, the bean beetle *Callosobruchus chinensis*. In this problem you will use BLAST to investigate these processes.

### 1. Part One:

Concisely define the following terms: Lateral Gene Transfer; retrotransposon; eukaryote; prokaryote; Polymerase Chain Reaction (PCR); Fluorescent In-situ Hybridization (FISH); endosymbiont; germ-line cells.

### 2. Part Two:

Blast the genome of *Drosophila melanogaster*'s *Wolbachia* symbiont against the *Drosophila melanogaster* genome and report your results. Blast the genome of *Drosophila melanogaster*'s *Wolbachia* symbiont against the *Drosophila ananassae* genome and report your results. The accession number of the *Wolbachia* symbiont we want you to use is NC\_002978. Be sure to report the accession numbers of the sequences you use or find.

### 3. Part Three:

Compare your findings. Can you come up with two separate hypotheses to explain your results? (Hint: think about how the two *Drosophila* species are related, and think about the Contamination Problem from your homework.)

## 1.7 Classic Alignments (5)

Use NCBI tools to retrieve the same alignments used in the classic paper:

- *Simian Sarcoma Virus onc Gene, v-sis, Is Derived from the Gene (or Genes) Encoding a Platelet-Derived Growth Factor*, R Doolittle.

Present the details of the tool settings. The document is available on <http://www.cs.brown.edu/courses/cs196-1/midterm.html>

## 1.8 Protein Structure (OPTIONAL 15)

In class we talked about protein folding and the Protein Data Bank (PDB). For this problem, we would like you to focus on the following four protein families (which can be found in the protein folding lecture):

- Flavodoxin-like fold Che-Y related
- Plastocyanin

- TIM Barrel
- Ferritin

First, we would like you to find three (3) protein examples of each family using the PDB. In your solution, include the name of the protein and a picture of its structure (structures of proteins from the same family should look similar).

Next, use NCBI tools to align the proteins and show that comparing pairs of proteins within each class gets larger scores than comparing proteins between different classes. Therefore, the scores of the alignments verify the classification. For example, if you compare two proteins from the TIM Barrel family you should always get a much higher score than comparing one of them with one from the plastocyanin family. Show five (5) such pairs of alignment examples and present your criteria used to align the sequences. You may find the tool “Align two sequences using BLAST (bl2seq)” under the category “Specialized BLAST” on the NCBI BLAST home page useful.

## 2 HOW NOT TO LIE WITH STATISTICS (10)

Over the course of the semester, you have been reading chapters from *How to Lie with Statistics*, to gain a better understanding of how statistics can be contorted to yield a bias towards a given result. Fabricate a “biological discovery” in both of the following categories which would look exceptional to a person not familiar with the biases that come with the biology of the domain, but which are not significant findings.

1. **Convergence or divergence (5)** There might be particular patterns of amino acids that are repeatedly selected for use as “turns.” Alternatively, there might be some combinations of amino acids that do not occur because of structural instability or steric problems. A further problem is that the 20 amino acids do not occur with equal frequencies. The three most frequent occurring amino acids are glycine, alanine, and leucine - account for a quarter of all residues. They occur four times as often as the least frequent amino acids - tryptophan, histidine, cysteine, and methionine.
2. **Identities (5)** The significance of percentage identity in an alignment between two proteins is very much a function of the lengths of the sequences being compared. So percentage identity by itself is not meaningful.

### 3 SUBSTITUTIONS MATRICES (15)

#### 3.1 BLOSUM-like Matrices (10)

As we discussed in class, when dealing with protein sequence alignment, substitutions are essential insight into biologically significant events. We introduced BLOSUM and PAM as approaches to calculating substitution matrices for proteins which take protein substitutions into consideration.

While substitutions usually refer to amino acid replacements, perhaps we can use the same methodology to consider how letters in the English language are replaced in different dialects. Consider the following excerpt from the Gold Bug story:

"Claws enoff, massa, and mouff too. I nabber did see sich a d--d bug --he kick and he bite ebervy ting what cum near him. Massa Will cotch him fuss, but had for to let him go gin mighty quick, I tell you --den was de time he must ha got de bite. I didn't like de look ob de bug mouff, myself, no how, so I wouldn't take hold ob him wid my finger, but I cotch him wid a piece ob paper dat I found. I rap him up in de paper and stuff piece ob it in he mouff --dat was de way."

Translation:

"Claws enough, master, and mouth too. I never did see such a darn bug --he kicked and he bit every thing that came near him. Master Will caught him first, but had for to let him go again mighty quick, I tell you --then was the time he must have got the bite. I didn't like the look of the bug's mouth, myself, no how, so I wouldn't take hold of him with my finger, but I caught him with a piece of paper that I found. I wrapped him up in the paper and stuffed a piece of it in his mouth --that was the way."

- Using these two versions of the gold bug story, we will be creating a substitution matrix following a method that is similar to (but not exactly) BLOSUM and PAM. Start by assuming that these two paragraphs have been aligned on white space, so that each word in the southern dialect is aligned with the corresponding word in "normal" English (if there are more/fewer characters, assume gaps have been inserted). Then using the beginning of each paragraph (all words up to "what cum near him"/"that came near him"), calculate by hand the frequencies of each letter being substituted with other letters.

For example, using "massa" and "master" we have:

m	a	s	s	a	-
↓	↓	↓	↓	↓	↓
m	a	s	t	e	r

So each letter has a probability of being substituted by other letters:

$$\begin{aligned} m &\rightarrow m(1.0) \\ a &\rightarrow a(0.5) \text{ or } e(0.5) \\ s &\rightarrow s(0.5) \text{ or } t(0.5) \\ - &\rightarrow r(1.0) \end{aligned}$$

Show the resulting substitution matrix.

### 3.2 Dayhoff Matrices (5)

What is the definition of PAM1? What was the role of the Markov Chain in the construction of the Dayhoff matrices?

## 4 GENOME ASSEMBLY (15)

### 4.1 Consensus Sequence (10)

Given a series of DNA sequences, determine the consensus DNA sequence (i.e. the DNA sequence from which all of the other DNA sequences can be derived with minimal changes).

1. Find the consensus DNA sequence from the following sequences:

```
ACACTACTAGGTAGACC
ACTCTACTAAGTCGACC
ACTGGAGCAAGTTGCGG
ATTCGACTTAGTCGCAG
ACTCGACTAAGTAGACG
ACACGACCAAGTCGACG
AGTGAAGTTAGCAGTCG
ACTCGACGAGATAGACG
ATTCGACTACGTAGACG
ACACGGGCAAGTAGCGG
```

2. Describe an algorithm to find all possible consensus sequences based on a number of sequences. When is there more than one?

### 4.2 Assembly Algorithms (5)

- What are the types of errors that make determining the reads overlap hard?
- How are mate-pairs useful in correcting misassembled contigs?
- How are mate-pairs useful in detecting order and orientation of contigs?

## 5 EXTRA CREDIT PROBLEMS

### 5.1 “The Adventure of the Dancing Men”

Read “The Adventure of the The Dancing Men” by Arthur Conan Doyle on <http://www.cs.brown.edu/courses/cs196-1/midterm.html>. This mystery has similarities with the Gold-Bug story and our analysis of it. In solving the mystery, similar methods are used with those used in Bioinformatics.

**Extra Credit (5 pts):** Present several analogies from the story that has Bioinformatics connotations relevant to the material in our class.

### 5.2 Tandem Repeats

In a given sequence a *tandem repeat* is a substring, which is immediately preceded by itself inside the sequence. Formally, for a sequence

$$X = x_1, x_2, x_3, \dots, x_n$$

a tandem repeat is a substring,  $w = x_i, \dots, x_j$  such that

$$x_k = x_{k+(j-i+1)} \quad i \leq k \leq j$$

**Extra Credit (5 pts):** Write a program that will locate a tandem repeat in a given sequence.

**Extra Credit (5 pts):** Revise your program so that it is  $O(n)$ .

**Extra Credit (5 pts):** Write a program to calculate the BLOSUM-like matrices from problem 3.1. Run it on the entire paragraph given and compare the resulting matrix with the one you found earlier.